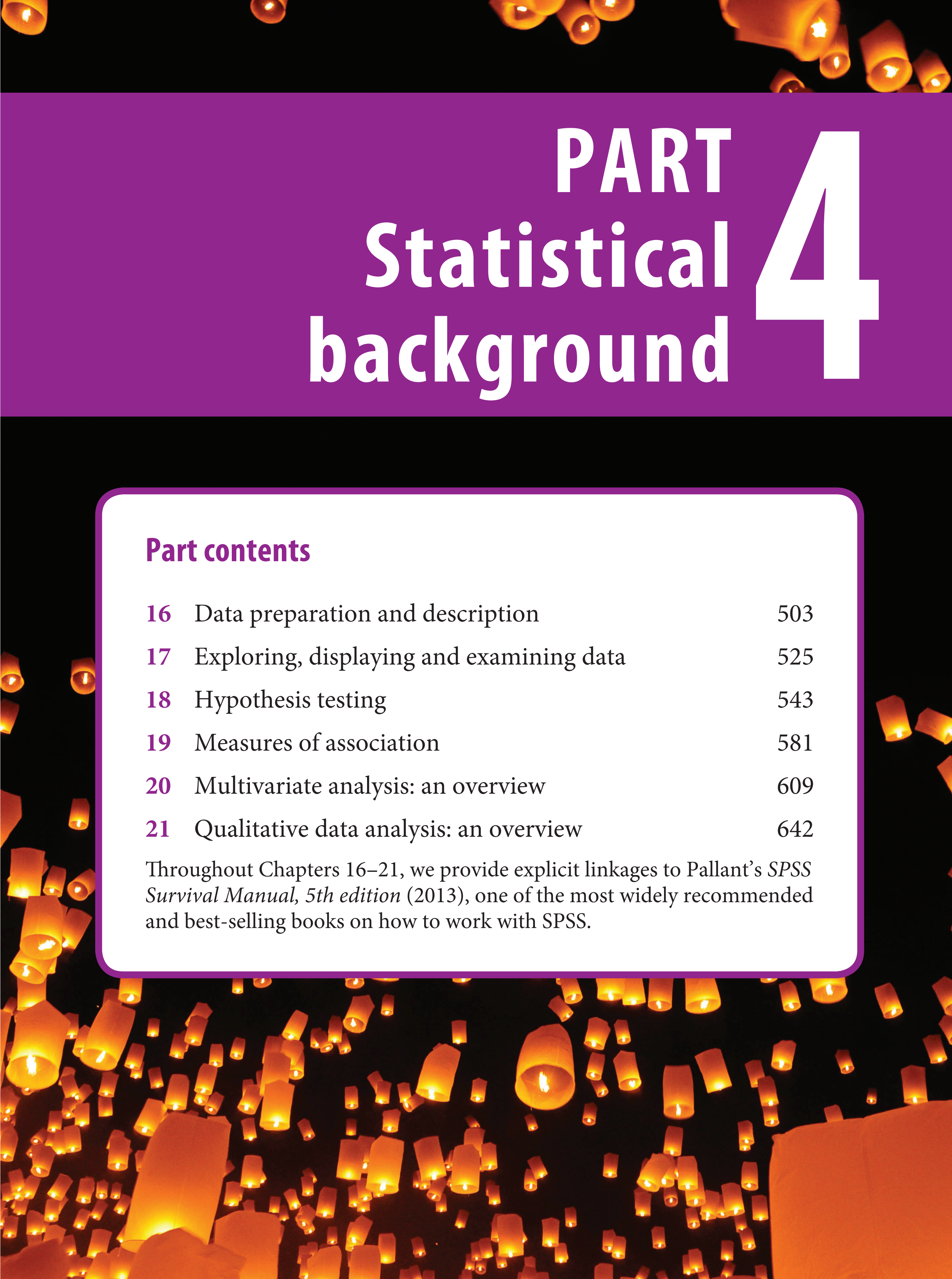


PART 4 Statistical background

Part contents

Throughout Chapters 16–21, we provide explicit linkages to Pallant's *SPSS Survival Manual, 5th edition* (2013), one of the most widely recommended and best-selling books on how to work with SPSS.

CHAPTER 16

Data preparation and description

Chapter contents

Learning objectives

When you have read this chapter, you should understand:

1 the importance of editing the collected raw data to detect errors and omissions

2 how coding is used to assign numbers to answers in order to classify responses

3 the use of content analysis to interpret and summarize open questions

4 problems and solutions for ‘don’t know’ responses

5 how to select descriptive statistics to summarize data and check for errors.

16.1 Introduction

Once the data begin to flow in, attention turns to data analysis. If the project has been organized and carried out correctly, the analysis planning is already done. Back in the research design stage, or at least by completion of the proposal or the pilot test, decisions should have been made about how to analyse the data. Unfortunately, many researchers wait until the analysis stage to decide what to do. This results in the late discovery that some data have not been collected, have been collected in the wrong form or will exhibit unanticipated characteristics.

This chapter addresses two topics. The first is data preparation, which includes editing, coding and data entry. These activities ensure the accuracy of the data and their conversion from raw form to reduced and classified forms that are more appropriate for analysis. Second, preparing a descriptive statistical summary is the preliminary step leading to an understanding of the collected data. This section is particularly valuable if your research objective is reporting or basic description, but descriptions also enrich any explanatory or predictive study, as they provide a thorough picture of the current state. We discuss the definitions and applications of descriptive statistics, specifically the characteristics of location, spread and shape that give us an insight into the distribution of respondent observations.

Before concluding this introduction, we must comment briefly on the reasons behind the provision of these chapters on the website. The website chapters were developed for and with graduate students and undergraduate students who had already followed an elementary statistics course. Without such grounding, some of the ideas presented may require remedial study.

16.2 Editing

The customary first step in analysis is to edit the raw data. Editing detects errors and omissions, corrects them where possible and certifies that minimum data quality standards have been achieved. The editor's purpose is to guarantee that the data are:

- accurate
- consistent with intent of the question and other information in the survey
- uniformly entered
- complete
- arranged to simplify coding and tabulation.

In the following question asked in a survey among employees of a manufacturing firm with different plant locations, one respondent checked two categories, indicating that he currently works in Bradford and Manchester.

Please indicate the location you currently work at:

- Bradford
- Hull
- Manchester (head office)
- Middleton
- Northampton

This double answer to the question can be true (e.g. a research and development (R&D) manager, who spends part of the week at head office and part of the week in the Bradford plant) or it can be a mistake during the interview (e.g. the respondent revised his or her answer but did not mark the revision clearly). The editor's responsibility is to decide which one of the responses is both consistent with the intent of the question or other information in the survey and most accurate for the individual respondent. If you want to avoid multiple answers to a single question, you need to reformulate the question, for example by adding in parentheses: (Please tick only one location. If you work in more than one location, tick the location where you spend most of your time.)

Field editing

Field editing review is a responsibility of the field supervisor, who is identical to the researcher in small projects, but can be a distinct person in (very) large projects. It, too, should be done soon after data have been gathered.

During the stress of data collection, the researcher often uses ad hoc abbreviations and special symbols, which are likely to be forgotten in a very short time, unless an adequate key is devised and followed. The legibility (or, more accurately, illegibility) of any notes taken during the research purpose can also be a problem. So, soon after the interview, experiment or observation, the researcher should review the reporting forms. When entry gaps are present from interviews, a call-back should be made rather than guessing what the respondent 'probably would have said'. Self-interviewing has no place in quality research.

In large projects, especially if you rely on professional interviewers from research agencies, the field supervisor has another important control function, namely to validate the field results. This normally means he or she will re-interview a percentage of the respondents, at least on some questions. Many commercial research firms will re-contact about 10 per cent of the respondents.

Central editing

At this point, the data should be subject to a thorough editing. For a small study, the use of a single editor produces maximum consistency. In large studies, the tasks may be broken down so that each editor can deal with one entire section. This approach will not identify inconsistencies between answers in different sections. However, this problem can be handled by identifying points of possible inconsistency and having one editor check for them specifically.

Sometimes it is obvious that an entry is incorrect, is entered in the wrong place or states time in months, say, when it was requested in weeks. When replies are clearly inappropriate or missing, the editor can sometimes detect the proper answer by reviewing the other information in the schedule. This practice, however, should be limited to those few cases where it is obvious what the correct answer is. It may be better to contact the respondent for correct information if time and budget allow.

Another alternative is for the editor to strike out the answer if it is clearly inappropriate. Here an editing entry of 'no answer' or 'unknown' is called for. This procedure, however, is not very useful if your sample size is small, as striking out an answer generates a missing value and often means that the observation cannot be used in the analyses that contain this variable. One approach to preserving the observation as part of the sample is to make an educated guess as to the likely answer and thus create a variable that includes all observations. Later on, if this variable has been included, at the very least it will tell you whether your guess has had a substantial effect on the results of the analysis. If not, your guess was accurate and unproblematic; if so, you might need to strike it out after all.

Another editing problem concerns faking an interview. This 'armchair interviewing' is difficult to spot, but the editor is in the best position to do so. One approach is to check responses to open-ended questions. These are the most difficult to fake. Distinctive response patterns in other questions will often emerge if faking is occurring. To uncover this, the editor must analyse the instruments used by each interviewer.

16.3 Useful rules for editing

Here are some useful rules to guide editors in their work:

- Familiarize yourself with the instructions given to interviewers and coders.
- Do not destroy, erase or make illegible any original entry by the interviewer; original entries should be crossed out with a single line so that they remain legible.
- Make all entries on an instrument in some distinctive colour and in a standardized form.
- Mark with your initials all answers changed or supplied.
- Write your initials and the date of editing on each instrument completed.

Coding

Coding involves assigning numbers or other symbols to answers so that the responses can be grouped into a limited number of classes or categories. The classifying of data into limited categories sacrifices some data detail but is necessary for efficient analysis. Instead of requesting the word 'male' or 'female' in response to a question that asks for the identification of one's gender, we could use the codes 'M' or 'F'. Normally this variable would be coded 0 for male and 1 for female. If we use M and F, or other letters, in combination with numbers and symbols, the code is alphanumeric. When numbers are used exclusively, the code is numeric. If you want to run quantitative analyses, you need to transform all alphanumeric codes to numbers, for example transform the 'M' and 'F' coding for gender to a dummy variable taking the values 0 for men and 1 for women. Alphanumeric variables that can take a lot of values are often coded using numbers too. There are even standardized coding schemes for common nominal alphanumeric variables, such as occupation or sector, where the numeric values represent a certain structure. Sectors are often coded according to SIC (standard industry classification) numbers, which have four digits at their lowest level and a common coding for occupations is ISCO88 (International Standard Classification of Occupations).

Coding helps the researcher to reduce several thousand replies to a few categories containing the critical information needed for analysis. In coding, categories are the partitioning of a set; and categorization is the process of using rules to partition a body of data.

Coding rules

Four rules guide the establishment of category sets. The categories should be:

- appropriate to the research problem and purpose
- exhaustive
- mutually exclusive
- derived from one classification principle.

Appropriateness

Categories must provide the best partitioning of data for testing hypotheses and showing relationships. Year-by-year age differences may be important to the question being researched. If so, wider age classifications hamper the analysis. If specific income, attitude or reason categories are critical to the testing relationship, then we must choose the best groupings. In particular, choose class boundaries that match those being used for comparisons. It is disheartening, late in a study, to discover that age, income or other frequency classes do not precisely match those of the data with which we wish to make a comparison.

Exhaustiveness

A large number of 'other' responses suggests that our classification set may be too limited. In such cases, we may not be tapping the full range of information in the data. Failure to present an adequate list of alternatives is especially damaging when multiple-choice questions are used. Any answer that is not specified in the set will surely be under-represented in the tally.

While the exhaustiveness requirement in a single category set may be obvious, a second aspect is less apparent. Does the one set of categories fully capture all the information in the data? For example, responses to an open-ended question about family economic prospects for the next year may initially be classified only in terms of being optimistic or pessimistic. It may also be enlightening to classify responses in terms of other concepts such as the precise focus of these expectations (income or jobs) and variations in responses between family heads and others in the family.

Mutual exclusivity

Another important rule is that category components should be mutually exclusive. This standard is met when a specific answer can be placed in one and only one cell in a category set. For example, in an occupation survey, the classifications may be (1) professional, (2) managerial, (3) sales, (4) clerical, (5) crafts, (6) operatives and (7) unemployed. Some respondents will think of themselves as being in more than one of these groups. The person who views selling as a profession and spends time supervising others may fit into three of these categories. A function of operational definitions is to provide categories that are composed of mutually exclusive elements. Here, operational definitions of the occupations to be classified in ‘professional’, ‘managerial’ and ‘sales’ should clarify the situation. The problem of how to handle an unemployed salesperson brings up a fourth rule of category design.

Single dimension

The need for a category set to follow a single classificatory principle means every class in the category set is defined in terms of one concept. Returning to the occupation survey example, the person in the study might be both a salesperson and unemployed. The ‘salesperson’ label expresses the concept occupation type; the response ‘unemployed’ is another dimension concerned with current employment status without regard to the respondent’s normal occupation. When a category set uses more than one dimension, it will normally not be mutually exclusive unless the cells in the set combine the dimensions (employed manager, unemployed manager, etc.). One way to handle the problem of multiple dimensions in the example above is to ask the respondent, first, about his or her employment status, and to follow this question about his occupation with a remark that indicates that those respondents who are currently unemployed are asked to answer this question with respect to the last job they held.

Codebook construction

A **codebook**, or coding scheme, contains each variable in the study and specifies the application of coding rules to the variable. It is used by the researcher or research staff as a guide to make data entry less prone to error and more efficient. It is also the definitive source for locating the positions of variables in the data file during analysis. In many statistical programs, the coding scheme is integral to the data file. Most codebooks – computerized or not – contain the question number, variable name, location of the variable’s code on the input medium, descriptors for the response options, and whether the variable is alpha or numeric. An example of a paper-based codebook is shown in Exhibit 16.1. When pilot testing has been conducted, there should be sufficient information about the variables to prepare a codebook. A preliminary codebook used with pilot data may reveal coding problems that will need to be corrected before the data for the final study are collected and processed.

*Exhibit 16.1 **Sample codebook of questionnaire items.***

Question	Variable number	Code description	Variable name
__________	1	Record number	RECNUM
__________	2	Respondent number	RESID
1	3	6-digit postcode	POSTC
		999999 = Missing	
2	4	4-digit birth year	BIRTH
		9999 = Missing	
3	5	Gender	GENDER
		1 = Male	
		2 = Female	
		9 = Missing	
4	6	Marital status	MARITAL
		1 = Married	
		2 = Widow(er)	
		3 = Divorced	
		4 = Separated	
		5 = Never married	
		9 = Missing	
5	7	Own–Rent	HOUSING
		1 = Own	
		2 = Rent	
		3 = Provided	
		9 = Missing	

▶

Exhibit 16.1 *Continued*

Question	Variable number	Code description	Variable name
6		Reason for purchase 1 = Mentioned 0 = Not mentioned	
	8	Bought home	HOME
	9	Birth of child	BIRTHCHD
	10	Death of a relative or friend	DEATH
	11	Promoted	PROMO
	12	Changed job/career	CHGJOB
	13	Paid college expenses	COLLEXP
	14	Acquired assets	ASSETS
	15	Retired	RETIRED
	16	Changed marital status	CHGMAR
	17	Started business	STARTBUS
	18	Expanded business	EXPBUS
	19	Parent's influence	PARENT
	20	Contacted by agent	AGENT
	21	Other	OTHER

Coding closed questions

The responses to closed questions include scaled items and others for which answers can be anticipated. When codes are established early in the research process, it is possible to pre-code the questionnaire. **Pre-coding** is particularly helpful for data entry because it makes the intermediate step of completing a coding sheet unnecessary. The data are accessible directly from the questionnaire. A respondent, interviewer, field supervisor or researcher (depending on the data-collection method) is able to assign an appropriate numerical response on the instrument by checking, circling or printing it in the proper coding location.

Exhibit 16.2 shows questions in the sample codebook. When pre-coding is used, editing may precede data processing. Note question 4, where the respondent may choose between five characteristics of marital status and enter the number of the item best representing present status in the coding portion of the questionnaire. This code is later transferred to an input medium for analysis.

'Don't know' responses

The **'don't know' (DK) response** presents special problems for data preparation. When the DK response group is small, it is not troublesome. But there are times when it is of major concern and it may even be the most frequent response received. Does this mean the question that elicited this response is useless? The answer is that it all depends. Most DK answers fall into two categories.[1] First, there is the legitimate DK response when the respondent does not know the answer. This response meets our research objectives; we expect DK responses and consider them to be useful.

In the second situation, a DK reply illustrates the researcher's failure to get the appropriate information. Consider the following illustrative questions:

1 Who received the 1992 Nobel Prize in economics?
2 Do you believe that the new government's fiscal policy is sound?
3 Do you like your present job?
4 Which of the various brands of chewing gum do you believe has the best quality?
5 How much do you spend each year for stationery?

Exhibit 16.2 ***Sample codebook of questionnaire items.***

1. What is the postcode of your residence? - - - - - - - - - - - - -
2. What is the year of your birth? 19 __________
3. Gender (1) Male
   (2) Female
   Indicate your choice by number → __________
4. What is your marital status?
   (1) Married
   (2) Widow(er)
   (3) Divorced
   (4) Separated
   (5) Never married
   Indicate your choice by number → __________
5. Do you own or rent your primary residence?
   (1) Own
   (2) Rent
   (3) Living quarters provided
   Indicate your choice by number → __________
6. What prompted you to purchase your most recent life insurance policy?
   ____
   ____
   ____
   ____
   ____
   ____
   ____

It is reasonable to expect that some legitimate DK responses will be made to each of these questions. In the first question, the respondents are asked for a level of information that they often will not have. There seems to be little reason to withhold a correct answer if known. Thus, most DK answers to this question should be considered as legitimate. A DK response to the second question presents a different problem. It is not immediately clear whether the respondent is ignorant of the government's fiscal policy or knows the policy but has not made a judgement about it. The researchers should have asked two questions. In the first, they would have determined the respondent's level of awareness of fiscal policy. If the interviewee passed the awareness test, then a second question would have secured judgement on fiscal policy.

In the remaining three questions, DK responses are more likely to be a failure of the questioning process, although some will surely be legitimate. The respondent may be reluctant to give the information. A DK response to question 3 may be a way of saying, 'I do not want to answer that question'. Question 4 might also elicit a DK response in which the reply translates to, 'This is too unimportant to talk about'. In question 5, the respondents are being asked to do some calculations about a topic to which they may attach little importance. Now the DK may mean, 'I do not want to do that work for something of so little consequence'.

Dealing with undesired DK responses

The best way to deal with undesired DK answers is to design better questions at the beginning. Researchers should identify the questions for which a DK response is unsatisfactory and design around it. Interviewers, however, often inherit this problem and must deal with it in the field. Several actions are then possible. First, good interviewer–respondent rapport will motivate respondents to provide more usable answers. When interviewers recognize an evasive DK response, they can repeat the question or probe for a more definite answer. The interviewer may also record verbatim any elaboration by the respondent and buck the problem on to the editor.

If the editor finds many undesired responses, little can be done unless the verbatim comments can be interpreted. Understanding the real meaning relies on clues from the respondent's answers to other questions. One way to do this is to estimate the allocation of DK answers from other data in the questionnaire. The pattern of responses may parallel income, education or experience levels. Suppose a question concerning whether employees like their present jobs elicits the answers in Exhibit 16.3. The correlation between years of service and the 'don't know' answers and the 'no' answers suggests that most of the 'don't knows' are disguised 'no' answers.

Exhibit 16.3 Handling 'don't know' responses.

Years of service	Do you like your present job?		
	Yes	No	Don't know
Less than 1 year	60%	10%	30%
1–3 years	50%	15%	35%
4 years or more	30%	20%	50%
N =	360	155	455

There are several ways to handle 'don't know' responses in the tabulations. If there are only a few, it does not make much difference how they are handled, but they will probably be kept as a separate category. If the DK response is legitimate, it should remain as a separate reply category. When we are not sure how to treat it, we should keep it as a separate reporting category and let the reader make the decision.

Another way to treat DK responses is to assume that they occur almost randomly. Using this approach, we distribute them among the other answers in the same ratio that the other answers occur. This assumes that those who reply 'don't know' are proportionally distributed among all the groups studied. This can be achieved either by prorating the DK responses or by excluding all DK replies from the tabulation. The latter approach is better since it does not inflate the actual number of other responses.

Coding open-ended questions

Given that your study has a quantitative nature, closed questions are favoured by researchers over open-ended questions for their efficiency and specificity. They are easier to measure, **record**, code and analyse. But there are situations where insufficient information or lack of a hypothesis prohibits preparing response categories in advance. Other reasons for using open-ended responses include the need to measure sensitive or disapproved behaviour, discover salience or encourage natural modes of expression.[2] However, analysing enormous volumes of open-ended questions has always been a nightmare for researchers. The variety of answers one may encounter is staggering. In Exhibit 16.2, question 6 illustrates the use of an open-ended question for which advance knowledge of response options was not available. The answer to 'What prompted you to purchase your most recent life insurance policy?' was to be filled in by the respondent as a short-answer essay. After preliminary evaluation, response categories (shown in the codebook example in Exhibit 16.1) were created for that item. Although most responses could be accounted for by the derived categories, an 'other' category was established to meet the coding rule of exhaustiveness.

SPSS reference

SPSS allows you to create codes and even a complete code book with a few commands. Pallant (2013) discusses how to code in Chapter 2. The commands to produce a codebook are presented in Chapter 6.

Using content analysis for open questions

In qualitative research, open questions are the dominant answer format and, as we have seen, quantitative research also uses open answers on occasion. Whether your research is qualitative or quantitative, coding open answers is useful in grasping the structure of the information collected. In both types of research, you are seeking to show which pieces of information go along with which other pieces or what differentiates between the information pieces. If the information in the answers to open questions cannot easily be transformed to numerical information, other methods have to be used. For further information on content analysis (see Chapter 10) and for information on how to code qualitative data (see Chapter 21).

Content analysis follows a systematic process, starting with the selection of a unitization scheme. The units may be syntactical, referential, propositional or thematic:

- Syntactical units are illustrated by words, which are the smallest and most reliable units.
- Referential units may be objects, events, persons, and so on, to which an expression refers. An advertiser may refer to a product as a 'classic', a 'power performer' or 'ranked first in safety' – each denoting the same object.
- Propositional units use several frameworks. One might show the relationships among the actor, the mode of acting and the object – for example, 'subscribers [actor] to this periodical save [mode of acting] €15 [object of the action] over the single issue rate'.
- Thematic units are higher-level abstractions inferred from their connection to a unique structure or pattern in the content. A response to a question about working conditions may reflect a temporal theme: the past ('how good things used to be here'), the present ('the need to talk with management now before production gets worse') or the future ('employee expectations to be involved in planning and goal setting').

Other aspects of the content analysis methodology include:

- election of a sampling plan
- development of recording and coding instructions
- data reduction
- inferences about the context
- statistical analysis.

Content analysis guards against selective perception of the content, provides for the rigorous application of reliability and validity criteria, and is amenable to computerization.

Let us look at an informal application of content analysis to a problematic open question. In this example, which we are processing without linguistics software technology, suppose employees in the assembly operation of a unionized manufacturing firm are asked, 'How can management–employee relations be improved?' A sample of the responses yields the following:

- Management should treat the workers with more respect.
- Managers should stop trying to speed up the assembly line.
- Working conditions in the shop are terrible. Managers should correct them.
- The foreman should be fired. He is unfair in his treatment of workers.
- Managers should form management–worker councils in the department to iron out problems and improve relations.
- Management should stop trying to undermine union leadership.
- Management should accept the union's latest proposals on new work rules.

The first step in analysis requires that the units developed reflect the objectives for which the data were collected. The research question is concerned with learning what the assemblers think is the locus of responsibility for improving company–employee relations. The categories selected are keywords and referential units. The first pass through the data produces a few general categories, as shown in Exhibit 16.4. These categories are mutually exclusive and contain only one concept dimension. The use of 'other' makes the category set exhaustive. If the sample suggested that many respondents identified the need for action by the public, government or regulatory bodies, then including all of them in 'other' would ignore much of the richness of the data.

Exhibit 16.4 ***Open question coding example (before revision).***

Question: 'How can management–employee relations be improved?'

Locus of responsibility	Mentioned	Not mentioned
A. Management	____________	____________
B. Union	____________	____________
C. Worker (other than union)	____________	____________
D. Joint management–union	____________	____________
E. Joint management–workers	____________	____________
F. Others	____________	____________

Since responses to this type of question often suggest specific actions, the second evaluation of the data uses propositional units. This identifies action objects and the actors previously discovered. If we used only the set of categories in Exhibit 16.3, the analysis would omit a considerable amount of information. The second analysis produces categories for action planning:

- human relations
- production processes
- working conditions
- other action areas
- no action area identified.

How can we categorize a response suggesting a combined management–production process? Exhibit 16.5 illustrates a combination of alternatives. By taking the categories of the first list with the action areas, it is possible to get an accurate frequency count of the joint classification possibilities for this question.

***Exhibit 16.5** Open question coding (after revision).*

Question: 'How can management–employee relations be improved?'		
Locus of responsibility		***Frequency*** **($n = 100$)**
A. Management		
	1. Human relations	15
	2. Production process	45
	3. Working conditions	25
	4. Other action areas	10
	5. No action area identified	5
B. Union		
	1. Human relations	10
	2. Production process	10
	3. Working conditions	65
	4. Other action areas	15
	5. No action area identified	0
C. Workers (other than union)		
D.		
E.		
F. Other		

Using available software, the researcher can spend much less time coding open-ended responses and capturing categories. Software also eliminates the high cost of sending responses to outside coding firms. What would take a coding staff several days can now be done in a few hours.

By applying statistical algorithms to create categories from open-ended survey responses, various programs allow you to import responses in a tab-delimited (ASCII) file format and then automatically penetrate the chaos of diverse answers by stemming, aliasing and excluding words that obscure important terms that are necessary to define meaningful categories.

- Stemming uses powerful linguistics technology to comb through responses, search for derivations of common root words and combine terms to create stemmed aliases (search, searching, searches).
- Aliasing combines synonyms (smart, wise, intelligent) into automatic aliases.
- Exclusion filters for trivial words (be, is, the, of) to create a list of included terms valuable for constructing meaningful categories.[3]

When using menu-driven programs, an autocategorization option creates manageable categories by clustering terms that occur together throughout the dataset. Then, with a few keystrokes, you can modify categorization parameters and refine your results. Once your categories are consistent with the research and investigative questions, you select what you want to export to a data file or in tab-delimited format. The output, in the form of tables and plots, serves as modules for your final report.

Research Methods in Real Life

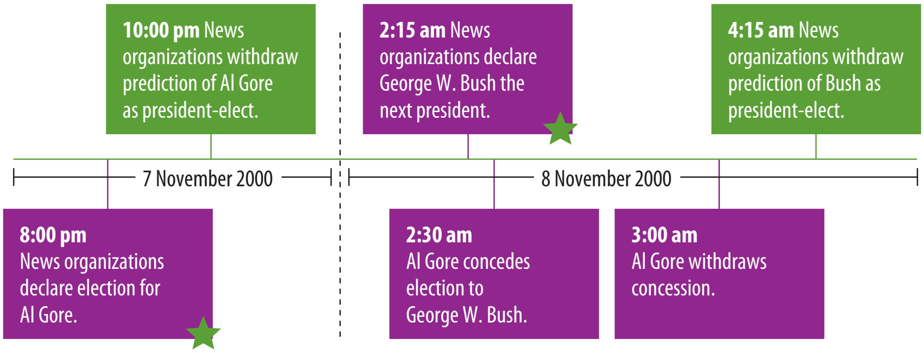
VNS: a black eye for research

A little before 8 pm on 7 November 2000, major news sources declared Al Gore the winner of Florida's electoral votes. Two hours later they pulled their prediction, reverting to 'too close to call'. At 2:15 am on 8 November, George W. Bush was declared the 43rd President and, soon after, Al Gore conceded the election. A month of Florida recounts gave the 2000 election its place in political, legal, news and research history.

Voter News Service (VNS), run by a consortium of ABC, NBC, CBA, CNN, FOX and the Associated Press, was responsible for the exit polls. Exit polls, regarded as reliable indicators of voting results, conduct intercept interviews with voters as they leave voting booths. Voters provide who they voted for, why, and extensive demographic and psychographic information. How could an established polling organization and internationally recognized news organizations' research fail so miserably?

The answer can be traced to the following factors: (i) the use of a biased sample, too many surveys completed in heavily Democratic precincts; (ii) mis-counted or mis-entered data in Duvall County (home of Jacksonville); and (iii) a rush to judgement for the news value of an early 'call' when, statistically, the margin between candidates was really too close to decide. Antitrust advocates claim that collusion versus competition resulted in poor methodology, and are seeking to break up VNS before another 'campaign fiasco' occurs (see Exhibit 16.6).

Exhibit 16.6 ***November 2000 US presidential election: timeline for poll-based actions.***



7 November 2000

8:00 pm News organizations declare election for Al Gore.

10:00 pm News organizations withdraw prediction of Al Gore as president-elect.

8 November 2000

2:15 am News organizations declare George W. Bush the next president.

2:30 am Al Gore concedes election to George W. Bush.

3:00 am Al Gore withdraws concession.

4:15 am News organizations withdraw prediction of Bush as president-elect.

References and further reading

'A bad day at the exit polls', Savanna Morning News, 9 November 2000 (savannahnow.com/110900/LOCsurvey.shtml).

'Antitrust group urges US to break up voter news service', Anti-Trust Institute, 27 November 2000 (www.antitrustinstitute.org/~antitrust/node/10346).

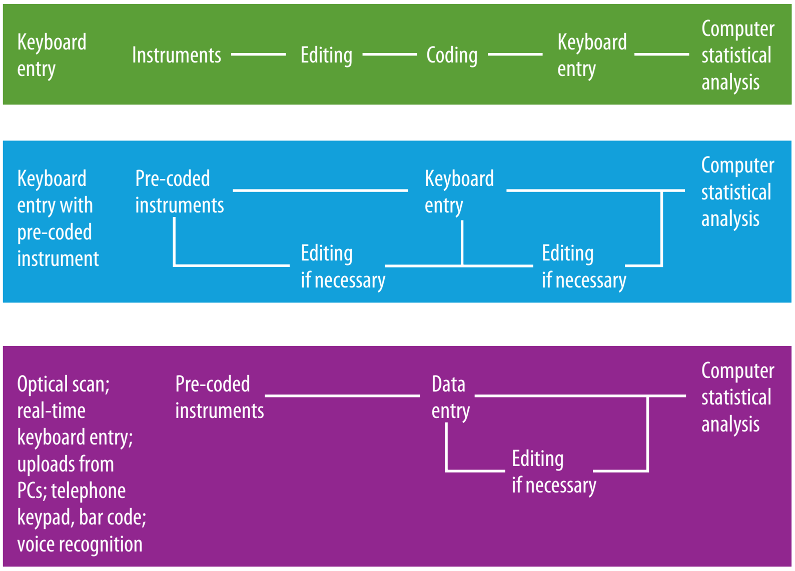
'Internet exit poll leaks election day', ABC News, 7 November 2000 (www.abcnews.go.com/Politics/story?id=122575).

Poll results (www.pollingreport.com/2000.htm).

16.4 Data entry

Researchers have profited from new and more efficient ways of speeding up the research process (see Exhibit 16.7). **Data entry** converts information gathered by secondary or primary methods to a medium for viewing and manipulation. Keyboarding remains a mainstay for researchers who need to create a data file immediately and store it in a minimal space on a variety of media.

Exhibit 16.7 Methods of data entry.


Keyboard entry
Instruments
Editing
Coding
Keyboard entry
Computer statistical analysis
Keyboard entry with pre-coded instrument
Pre-coded instruments
Keyboard entry
Editing if necessary
Editing if necessary
Computer statistical analysis
Optical scan; real-time keyboard entry; uploads from PCs; telephone keypad, bar code; voice recognition
Pre-coded instruments
Data entry
Editing if necessary
Computer statistical analysis

If you use a PC image scanner, you probably are familiar with **optical character recognition** programs that transfer printed text into computer files in order to edit and use it without retyping. There are other, related applications. **Optical scanning** instruments – the choice of testing services – are efficient for researchers. Examinees darken small circles, ellipses or sets of parallel lines to choose a test answer. Optical scanners process the marked-sensed questionnaires and store the answers in a file. This method, most often associated with standardized and pre-printed forms, has been adopted by designers for data entry and pre-processing. It reduces the number of times data are handled, thereby reducing the number of errors that are introduced.

In a far more flexible format, **optical mark recognition (OMR)** uses a spreadsheet-style interface to read and process user-created forms. The primary advantages are:

- speed – about 10 times faster than manual keying
- accuracy – error reduction from keying or stray marks
- cost – savings on data entry, form design and reproduction
- convenience – makes viewing the data easy, and provides charts and reports.

The researcher creates OMR forms with a word processor, page layout program or survey design package and includes bubbles, fill-in marks, checkboxes, bar codes and image fields. Data collected from questionnaires, reader reply cards, tests, interviews, observation forms or checklists – virtually any paper-based medium – can be processed this way. A photocopier, laser printer or local print shop then duplicates the forms. Special papers, drop-out inks or the pencils associated with traditional OMR are not needed since the respondent may use markers, pens, pencils or even crayons with the plain paper forms.

Most OMR products have 'trainable' software, so your questionnaire need not conform to a particular design. Training involves creating a template by scanning in your form, displaying the image and then, with a mouse, dragging boxes around the marking areas to define how each mark should be translated.[4] Once the template is created, a flatbed scanner can be used to read the data. For projects with hundreds or thousands of pages, a high-speed, sheet-fed scanner is essential. Scanned images of the collected data may be edited, saved and processed with tabulation tools internal to the software. Many researchers export the results in multiple formats to statistical, database and spreadsheet applications for more sophisticated analysis.

In addition to reading pre-scanned forms and questionnaire faxes, one program allows you to convert your paper-based OMR surveys into online surveys for the Internet or an intranet and then merge the results.[5]

The **bar code** is a well-known example of how to facilitate OMR. The first actual bar code systems went into a General Motors plant to monitor the production and distribution of automobile axles and to the General Trading Company for directing shipments to the proper loading bays. After a 1970s study by McKinsey & Company that predicted unprecedented savings in the grocery industry, the Kroger grocery chain pilot-tested a production system and bar codes became ubiquitous in that industry.[6]

The bar code is split into two halves of six digits each. It is used in numerous applications: point-of-sale terminals, hospital patient ID bracelets, inventory control, product and brand tracking, promotional technique evaluation, shipping cartons, marathon runners, at rental car locations to speed the return of cars and generate invoices and to track insects' mating habits. The codes appear on business documents, truck parts and timber in lumberyards. Federal Express shipping labels use a code called Codabar. Improvements to the basic universal product code (UPC) include the European Article Numbering (EAN) system, which has an extra pair of digits and may become the world's most widely used system.

Bar code technology is used to simplify the interviewer's role as a data recorder. Instead of writing (or typing) information about the respondents and their answers by hand, the interviewer can pass a bar code wand over the appropriate codes. The data are recorded in a small, lightweight unit for translation later. In the large-scale processing project Census 2000, the US Census Data Capture Center used bar codes to identify residents.

Other techniques include direct response entry, of which voting procedures used in several states are an example. With a specially prepared punch card, citizens cast their votes by pressing a pen-shaped instrument against the card next to the preferred candidate. This opens a small hole in a specific column and row of the card. The cards are collected and placed directly into a card reader. This method also removes the coding and entry steps. One problem with these punch cards is the respondent's accuracy in punching the hole. In the US presidential election 2000, punch cards used for voting in Florida generated many invalid votes, as vote counters were unable to decide whether a punched hole was a vote for Bush or Gore.

The declining cost of technology has allowed most researchers access to desktop or portable computers or networks to larger computers. This technology enables computer-assisted telephone or personal interviews to be completed with answers entered directly for processing, eliminating intermediate steps and errors. With a built-in communications modem or cellular link, their files can be sent directly to another computer in the field or to a remote site. Similarly, in the growing field of web-based surveys, respondents enter their answers directly into a database. **Computer-assisted interviewing** and **web-based surveys** economize on the data entry, because in these kinds of survey the respondents rather than the interviewers enter the data (the respondent's answers) into the computer. Still, researchers need to inspect the data that have been entered and stored.

The increase in computerized random-digit dialling has encouraged other data collection innovations. **Voice recognition** and response systems, while still far from mature, are providing some interesting alternatives for the telephone interviewer. Such systems can be used with software that is programmed to call specific three-digit prefixes and randomly generated four-digit numbers, reaching a random sample within a set geographical area. On getting a voice response, the computer branches into a questionnaire routine. Currently, the systems are programmed to record the verbal answers, but voice recognition is improving quickly enough so that these systems will soon translate voice responses into data files. Telephone keypad response is another capability made possible by computers linked to telephone lines. Using the telephone keypad (touch tone), the respondent answers questions by pressing the appropriate number. The computer captures the data by 'listening', decoding the tone's electrical signal, and storing the numeric or alphabetic answer in a data file.

Even with these time reductions between data collection and analysis, continuing innovations in multimedia technology are being developed by the personal computer business. The capability to integrate visual images, streaming video, audio and data may soon replace video equipment as the preferred method for recording an experiment, interview or focus group. A copy of the response data could be extracted for data analysis, but the audio and visual images would remain intact for later evaluation.

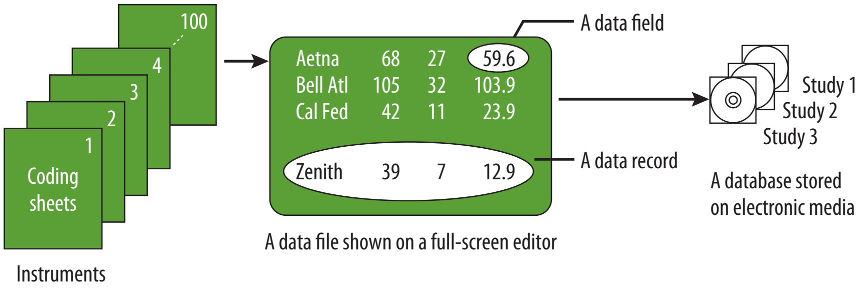
Other techniques on the horizon will continue to improve research efficiency and effectiveness. Although technology will never replace researcher judgement, it can reduce data-handling errors, decrease time between data collection and analysis, and help provide more usable information.

Data entry formats

A full-screen editor, where an entire **data file** can be edited or browsed, is a viable means of data entry for statistical packages like SPSS, STATA or SAS. The same software makes accessing data from databases, spreadsheets, data warehouses or data marts effortless.

For large projects, database programs serve as valuable data entry devices. A **database** is a collection of data organized for computerized retrieval. Programs allow users to define **data fields** and link files so that storage, retrieval and updating are simplified. The relationship between data fields, records, files and databases is illustrated in Exhibit 16.8. A company's personnel records serve as an example of a database. Employee information may be kept in several files: salary and position, education, benefits, and home and family. The data are separated so that authorized people can see only those parts pertinent to their needs. However, the files may be linked so that when, say, a woman changes her name, the change is entered once and all the files are updated.

Exhibit 16.8 ***Data fields, records, files and databases.***



Source: company data; public domain

Note: Data fields represent single elements of information (e.g. an answer to a question, a description, a number, a statement). Data fields can contain numeric, alphabetic or symbolic information. A record is a set of data fields that are related. Records are the rows of a data file or spreadsheet program worksheet. Data files are sets of records that are grouped together for storage on disks, tapes, CDs or hard drives.

Databases are made up of one or more data fields that are interrelated.

Researchers consider database entry when they have large amounts of potentially linked data that will be retrieved and tabulated in different ways over time. Another application of a database program is as a 'front-end' entry mechanism. A telephone interviewer may ask the question, 'How many children live in your home?' The computer's software has been programmed to accept any answer between 0 and 20. If a 'P' is accidentally struck, the program will not accept the answer and will return the interviewer to the question. With a pre-coded online instrument, much of the editing needed previously is done by the program. In addition, the program can be set for automatic conditional branching. In the example, an answer of 1 or greater causes the program to prompt the questioner to ask the ages of the children. A 0 causes the age question to be automatically skipped. Although this option is available whenever interactive computing is used, front-end processing is typically done within the database design. The database will then store the data into a set of linked files that allow the data to be easily sorted. **Descriptive statistics** and tables are readily generated from within the database.

Spreadsheets are a specialized type of database. For data that need organizing, tabulating and simple statistics, spreadsheets provide an easy-to-learn mechanism. They also offer some database management, graphics and presentation capabilities. Data entry on a spreadsheet uses numbered rows and letter columns with a matrix of thousands of cells into which an entry may be placed. Spreadsheets allow you to type numbers, formulas and text into appropriate cells. Many statistics programs for spreadsheet format are shown in Exhibit 16.9. This is a convenient and flexible means for entering and viewing the data. PCs also have charting and graphics applications and have data editors similar to Microsoft Excel™.

Exhibit 16.9 ***Data entry using spreadsheets.***

	A	B	C	D	E	F	G	H
1	City	Population	Men	Women	Density	Demographic pressure	Non-western foreigners	Largest party local elections
2	Beek	17 145	8 484	8 661	817	60	455	3
3	Born	14 630	7 303	7 327	654	57,40000153	290	3
4	Brunssum	30 277	14 798	15 479	1757	64.5	970	3
5	Eijsden	12 038	6 006	6 032	609	60,70000076	120	3
6	Goleen	34 117	16 759	17 359	1 383	63	2 115	3
7	Gulpen-Wittem	15 537	7 772	7 765	212	57,59999847	205	3
8	Heerlen	95 367	46 853	48 514	2 103	61,70000076	5 270	3
9	Kerkrade	51 762	25 416	26 346	2 358	60	1 475	3
10	Landgraaf	41 411	20 492	20 919	1 678	57,40000153	785	3
11	Maastricht	121 479	58 376	63 103	2 129	56.5	6 380	1
12	Margraten	13 780	7 025	6 755	239	60,70000076	170	1
13	Meerssen	20 308	10 041	10 267	740	58.5	300	1
14	Nuth	16 642	8 304	8 338	502	57,79999924	510	3
15	Ondarbanken	8 506	4 250	4 256	401	59,70000076	100	3
16	Schinnen	13 759	6 954	6 805	572	58,20000076	300	3
17	Simpelveld	11 617	5 792	5 825	725	57,90000153	105	3
18	Sittard	49 524	24 490	25 034	1589	57,90000153	2 075	3
19	Stein	26 303	13 206	13 097	1233	55,40000153	425	2
20	Susteren	13 033	6 526	6 507	439	56,70000076	175	1
21	Vaals	10 981	5 435	5 546	460	59,09999847	210	3
22	Valkenburg	17 908	8 855	9 053	488	59,20000076	325	3

Source: Statistics Netherlands; Statline.

SPSS reference

SPSS supports data entry with many utilities that allows you to modify and manipulate variables, single observations and even full datasets. Pallant (2013) discusses these commands in Chapter 4 and some checks SPSS can perform in Chapter 5.

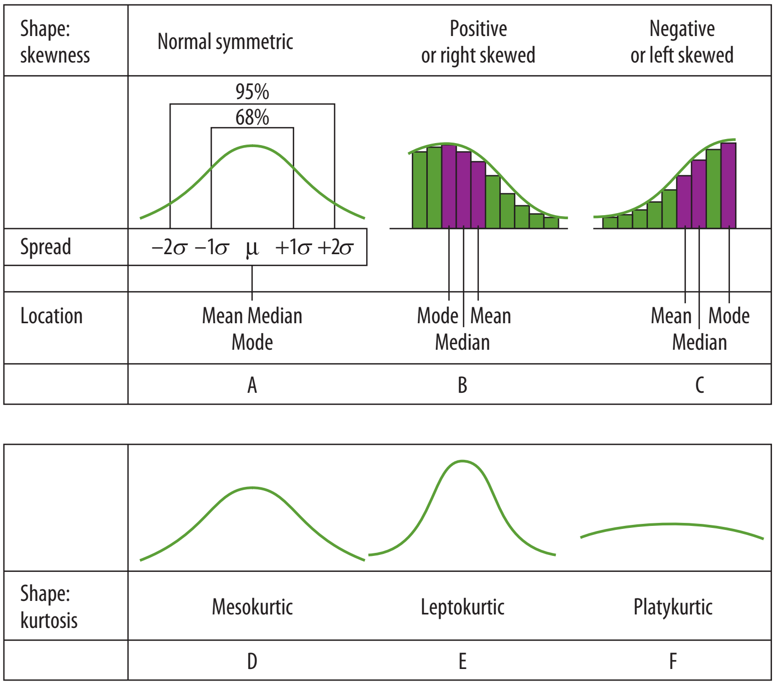
A **data warehouse** organizes large volumes of data into categories to facilitate retrieval, interpretation and sorting by end-users. Data warehouse tools are of two types: (i) transformation and cleansing, and (ii) end-user access tools. Both primary and secondary data can be warehoused. With survey responses, for example, one transformation process involves binning the classification variables. A respondent's occupation could have a value from 1 to 20. If each occupational code was allowed to remain a discrete value, the information might be lost because of sparse data in that category. Binned responses provide more data points and increase the information within each essential category. Data quality is a problem during transformation and cleansing. To avoid useless information, the data should have few missing values. The key is to monitor data continually as it is added to the data warehouse, using the exploratory techniques described in Chapter 17.

16.5 Descriptive statistical summaries

In the first part of this chapter, we discussed how responses are coded and entered. Creating numerical summaries of this process provides valuable insights into its effectiveness. For example, **missing data**, information that is missing about a respondent or case for which other information is present, may be detected. Miscoded, out-of-range data, extreme values and other problems also may be rectified after a preliminary look at the dataset.[7] To understand each variable's characteristics, first consider the type of scale on which it was measured. With nominal measurements (e.g. company classifications like high-technology, financial and retailing), each category is represented by a numerical code that refers back to a verbal description of the category. With ordinal data, the item's rank, reflecting a position in the range from the lowest to the highest, is entered. The same is true with interval-ratio scores. When these data are tabulated, they may be arrayed from the lowest to the highest scores on their scales. Together with the frequency of occurrence, the observations form a distribution of values.

Many variables of interest have distributions that approximate to a **standard normal distribution**. This distribution, shown in Part A of Exhibit 16.10, is the most significant theoretical distribution in statistics. It is a standard of comparison for describing distributions of sample data and is used with inferential statistics that assume normally distributed variables.

Exhibit 16.10 Characteristics of distributions.



Look at Exhibit 16.11 and examine the sample distribution of variables from a dataset on responses of employees in one department to a job satisfaction survey. These data were collected on a five-point interval scale. There are no missing data in variable 'education' (1A), although it is apparent that a range of 6 and a maximum value of 7 invalidate the calculated mean or average score. The variables 'functional level' (1B) and 'seniority' (2B) have one case missing but values that are within range. Variable 'job satisfaction' (2A) is missing four cases, or 27 per cent of its data points. The last variable, 'loyalty' (2C), has a range of 6, one missing value and three values coded as '9'. A '9' is often used as a DK or missing value code when the scale has a range less than nine points. In this case both blanks and 9s are present – a coding concern.

Notice that the fifth respondent answered only two of the five questions, and the second respondent had two miscoded answers and one missing value. Finally, using descriptive indexes of shape, discussed later in this section, you can find three variables that depart from the symmetry of the normal distribution. They are skewed (or pulled) to the left by a disproportionately small number of 1s and 2s. One variable's distribution is peaked beyond normal dimensions.

We have only used the minimum and maximum values, the range and mean, and have already discovered errors in coding, problems with respondent answer patterns and missing cases.

Now let us look at some other descriptive tools for this purpose. The characteristics of location, spread and shape are helpful initial tools for cleaning the data, discovering problems and summarizing distributions. Their definitions, applications and formulas fall under the heading of descriptive statistics. Although the definitions will be familiar to most readers, we take the following perspective on the characteristics of distributions:

- A distribution's shape is just as consequential as its location and spread.
- The choice of summary statistics to describe a single variable is contingent on the appropriateness of those statistics for the shape of the distribution.
- Visual representations are ultimately superior to numerical ones for discovering a distribution's shape and should be used before selecting remedies to correct anomalies in the data.[8]

***Exhibit 16.11* Dataset example: missing and out-of-range data.**

Case	1A	1B	2A	2B	2C
1	5.0	5.0	5.0	5.0	9.0
2	7.0	3.0		4.0	9.0
3	5.0	5.0	5.0	5.0	5.0
4	5.0	5.0	4.0		
5	1.0			2.0	
6	5.0	5.0	5.0	5.0	9.0
7	5.0	5.0	5.0	5.0	5.0
8	4.0	3.0	3.0	3.0	3.0
9	4.0	4.0	5.0	5.0	5.0
10	4.0	5.0		4.0	5.0
11	2.0	5.0	4.0	4.0	5.0
12	6.0	4.0	3.0	3.0	4.0
13	5.0	5.0		3.0	5.0
14	5.0	5.0	5.0	5.0	5.0
15	5.0	4.0	5.0	5.0	4.0
Valid	15	14	11	14	13
Missing	0	1	4	1	2
Mean	4.53	4.50	4.45	4.14	5.61
Range	6	2	2	3	6
Minimum	1	3	3	2	3
Maximum	7	5	5	5	9

Legend: 1A: education; 1B: functional level; 2A: job satisfaction; 2B: seniority; 2C: loyalty.

Measures of location

Summarizing the information from our collected data often requires the description of 'typical' values. The data in Exhibit 16.9 are a subset of a database with basic demographic information of Dutch cities. Suppose we want to know the typical score for a city's population density. We might define typical as the average response (mean); the middle value, when the distribution is sorted from lowest to highest (median); or the most frequently occurring value (mode). The common **measures of location**, often called **central tendency** or centre, include the **mean**, **median** and **mode**.

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

For the density variable in Exhibit 16.9, the distribution of responses is: 817, 654, 1757, 609, 1383, 212, 2103, 2358, 1678, 2129, 239, 749, 502, 401, 572, 725, 1589, 1233, 439, 460, 488, 414. The arithmetic average, or mean, of the 22 values is: (817+654+1757+609+1383+212+2103+2358+1678+2129+239+749+502+401+572+725+1589+1233+439+460+488)/21≈1005 (or an average population density of 1005). Note that three of the 21 cities have a density of more than 2000.

The median is the midpoint of the distribution. Half of the observations in the distribution fall above and the other half fall below the median. When the distribution has an even number of observations, the median is the average of the two middle scores. The median is the most appropriate locator of centre for ordinal data and has resistance to extreme scores, thereby making it a preferred measure for interval-ratio data when their distributions are not normal. The median is sometimes symbolized by M or mdn.

From the sample distribution for density variable, the median of the 21 values is 725. The distribution is ordered: 212, 239, 401, 439, 460, 488, 502, 572, 609, 654, 725, 749, 817, 1233, 1383, 1589, 1678, 1757, 2103, 2129, 2358.

The mode is the most frequently occurring value. When there is more than one score that has the highest yet equal frequency, the distribution is bimodal or multi-modal. When every score has an equal number of observations, there is no mode. The mode is the location measure for nominal data and a point of reference along with the median and mean for examining spread and shape. In our example, the most frequently occurring value for the variable 'largest party in local elections' is 3.

Measures of spread

The common **measures of spread**, alternatively referred to as dispersion or variability, are the variance, standard deviation, range, interquartile range and quartile deviation. They describe how scores cluster or scatter in a distribution.

The variance is the average of the squared deviation scores from the distribution's mean. It is a measure of score dispersion about the mean. If all the scores are identical, the variance is 0. The greater the dispersion of scores, the greater the variance. Both the variance and the standard deviation are used with interval-ratio data. The symbol for the sample variance is s^2 and the population variance is the Greek letter sigma, squared (σ^2). The variance is computed by summing the squared distance from the mean for all cases and dividing the sum by the total number of cases minus one.

$$S^2 = \sum_{i=1}^{n} \frac{(X - \bar{X})^2}{n - 1}$$

For the density variable, we would compute the variance as:

$$s^2 = (817 - 978)^2 + (654 - 978)^2 + (1757 - 978)^2 \cdots + (488 - 978)^2/21 = 460941$$

The **standard deviation** summarizes how far away from the average the data values typically are. It is perhaps the most frequently used measure of spread because it improves interpretability by removing the variance's square and expressing deviations in their original units (e.g. revenues in euros, not euros squared). It is also an important concept for descriptive statistics because it reveals the amount of **variability** of individuals within the dataset. Like the mean, the standard deviation is affected by extreme scores. The symbol for the sample standard deviation is s and a population standard deviation is σ. Alternatively, it is labelled '*stddev*'.

$$stddev = \sqrt{s^2}$$

The standard deviation for the density variable in our example is 679. The **range** is the difference between the largest and smallest score in the distribution. The density variable has a range of 2146 (2358 – 212). Unlike the standard deviation, it is computed from only the minimum and maximum scores; thus, it is a very rough measure of spread. With the range as a point of comparison, it is possible to get an idea of the homogeneity (small stddev) or heterogeneity (large stddev) of the distribution. For homogeneous distribution, the ratio of the range to the standard deviation should be between 2 and 6. In the density example, the ratio is 2146/674 = 3.18. A number above 6 would indicate a high degree of heterogeneity. The range provides useful but limited information for all data. It is mandatory for ordinal data.

The **interquartile range (IQR)** is the difference between the first and third quartiles of the distribution. It is also called the midspread. Ordinal or ranked data use this measure in conjunction with the median. It is also used with interval-ratio data when asymmetrical distributions are suspected or for exploratory analysis. Recall the following relationships. The minimum value of the distribution is the 0th percentile; the maximum, the 100th percentile. The first quartile (Q_1) is the 25th percentile; the median, or Q_2, is the 50th percentile. The third quartile (Q_3) is the 75th percentile.

The **quartile deviation**, or semi-interquartile range, is expressed as

$$Q = \frac{Q_3 - Q_1}{2}$$

The quartile deviation is always used with the median for ordinal data. It is helpful for interval-ratio data of a skewed nature. In a normal distribution, the median plus one quartile deviation (Q) on either side encompasses 50 per cent of the observations. Eight Qs cover approximately the range. Q's relationship with the standard deviation is constant (Q = .6745*s*) when scores are normally distributed. For the density variable, Q is: 1537 – 467/2 = 535.

Measures of shape

The measures of shape, skewness and kurtosis describe departures from the symmetry of a distribution and its relative flatness (or peakedness), respectively. They are related to statistics known as 'moments', which use deviation scores ($X - \bar{X}$). The **variance**, for example, is a second power moment. The **measures of shape** use third and fourth power deviations for their computations and are often difficult to interpret when extreme scores are in the distribution. Generally, shape is best communicated through visual displays. From a practical standpoint, the calculation of skewness and kurtosis is best done with spreadsheet or statistics software.

Skewness is a measure of a distribution's deviation from symmetry. In a symmetrical distribution, the mean, median and mode are in the same location. A distribution that has cases stretching towards one tail or the other is called skewed. As shown in Exhibit 16.10, when the tail stretches to the left, to smaller values, it is negatively skewed. Scores stretching towards the right, towards larger values, skew the distribution positively. Note the relationship between the mean, median and mode in asymmetrical distributions. The mean and standard deviation are called dimensional measures. That is, they are expressed in the same units as the measured quantities. In contrast, skewness is considered a non-dimensional measure because it is an index that only characterizes the shape of the distribution. The symbol for skewness is *sk*.

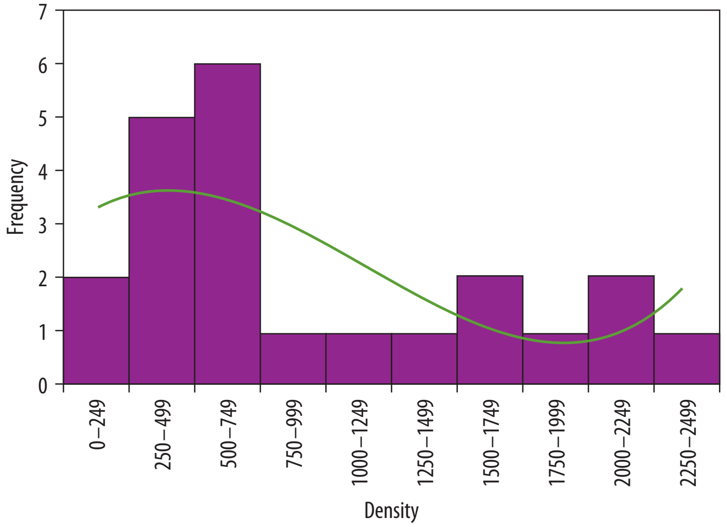
$$sk = \frac{n}{(n-1)(n-2)} \mathrm{S} \left(\frac{x_i - \bar{x}}{x} \right)^3$$

where *s* is the sample standard deviation (the unbiased estimate of sigma).

When a distribution approaches symmetry, *sk* is approximately 0. With a positive skew, *sk* will be a positive number; with negative skew, *sk* will be a negative number. The calculation of skewness for our sample density data produces an index of 0.744 and reveals a positive skew (see Exhibit 16.12).

As illustrated in the lower portion of Exhibit 16.10, kurtosis is a measure of a distribution's peakedness (or flatness). It is also a non-dimensional index. Distributions that have scores that cluster heavily or pile up in the centre (along with more observations than normal in the extreme tails) are peaked or leptokurtic. Flat distributions,

Exhibit 16.12 ***Shape characteristics in spreadsheet variable density.***



with scores more evenly distributed and tails fatter than a normal distribution, are called platykurtic. Intermediate or mesokurtic distributions are neither too peaked nor too flat. The symbol for kurtosis is *ku*.

$$ku = \left\{ \frac{b(n+1)}{(n-1)(n-2)(n-3)} S \left(\frac{x_i - \bar{x}}{x} \right)^4 \right\} - \frac{3(n-1)^3}{(n-2)(n-3)}$$

where *s* is the sample standard deviation (the unbiased estimate of sigma).

The value of *ku* for a normal or mesokurtic distribution is close to 0. A leptokurtic distribution will have a positive value and the platykurtic distribution will be negative. As with skewness, the larger the absolute value of the index, the more extreme is the characteristic. In the population density example, the kurtosis is calculated as –0.852, which suggests a slight deviation from a normally shaped curve with peaking contributed by greater frequency of values in the range of 250 to 750 million (see Exhibit 16.12).

SPSS reference

The SPSS commands to get these descriptive information on variables are discussed in Pallant (2013) Chapter 6.

Summary

1 The first step in data preparation is to edit the collected raw data to detect errors and omissions that would compromise quality standards. The editor is responsible for making sure that the data are accurate, consistent with other data, uniformly entered and ready for coding. In survey work, it is common to use both field and central editing.

2 Coding is the process of assigning numbers and other symbols to answers so that we can classify the responses into categories. Categories should be appropriate to the research problem, exhaustive of the data, mutually exclusive and one-dimensional. The reduction of information through coding requires the researcher to design category sets carefully, using as much of the data as possible. Codebooks are guides to reduce data entry error and serve as a compendium of variable locations and other information for the analysis stage.

3 Closed questions include scaled items and other items for which answers are anticipated. Pre-coding of closed items avoids tedious completion of coding sheets for each response. Open questions are more difficult to code since answers are not prepared in advance, but they do encourage disclosure of complete information. A systematic method for analysing open questions is called content analysis. It uses pre-selected sampling units to produce frequency counts and other insights into data patterns.

4 'Don't know' responses are evaluated in light of the question's nature and the respondent. While many DKs are legitimate, some result from questions that are ambiguous or from an interviewing situation that is not motivating. It is better to report DKs as a separate category unless there are compelling reasons to treat them otherwise.

5 Data entry is accomplished by keyboard entry from pre-coded instruments, optical scanning, real-time keyboarding, telephone pad data entry, bar codes, voice recognition, OMR, and data transfers from electronic notebooks and laptop computers. Database programs, spreadsheets and editors in statistical software programs offer flexibility for entering, manipulating and transferring data for analysis, warehousing and mining.

6 The objective of descriptive statistical analysis is to develop sufficient knowledge to describe a body of data. This is accomplished by understanding the data levels for the measurements we choose, their distributions, and characteristics of location, spread and shape. The discovery of miscoded values, missing data and other problems in the dataset is enhanced with descriptive statistics.

Discussion questions

Terms in review

1 Define or explain:

  a coding rules

  b spreadsheet data entry

  c bar codes

  d pre-coded instruments

  e measures of shape

  f content analysis

  g missing values

  h optical mark recognition

  i measures of spread.

2 How should the researcher handle 'don't know' responses?

3 Why is the standard deviation a more useful statistic than the variance?

Making research decisions

4 A problem facing shoe store managers is that many shoes must eventually be sold at markdown prices. This prompts us to conduct a mail survey of shoe store managers in which we ask, 'What methods have you found most successful for reducing the problem of high markdowns?' We are interested in extracting as much information as possible from these answers to understand better the full range of strategies that store managers use. Establish what you think are category sets to code 500 responses similar to the 14 given below. Try to develop an integrated set of categories that reflects your theory of markdown management. After developing the set, use it to code the 14 responses.

  a Have not found the answer. As long as we buy style shoes, we will have markdowns. We use PMs on slow merchandise, but it does not eliminate markdowns. (PM stands for 'push-money' – special item bonuses for selling a particular style of shoe.)

  b Using PMs before too old. Also reducing price during season. Holding meetings with salespeople indicating which shoes to push.

  c By putting PMs on any slow-selling items and promoting same. More careful check of shoes purchased.

  d Keep a close watch on your stock and mark down when you have to – that is, rather than wait, take a small markdown on a shoe that is not moving at the time.

  e Using the PM method.

  f Less advance buying – more dependence on in-stock shoes.

  g Sales – catch bad guys before it is too late and close out.

  h Buy as much good merchandise as you can at special prices to help make up some markdowns.

  i Reducing opening buys and depending on fill-in service. PMs for salespeople.

  j Buy more frequently, better buying, PMs on slow-moving merchandise.

  k Careful buying at lowest prices. Cash on the buying line. Buying closeouts, FDs (factory discontinued), overstock, 'cancellations'.

  l By buying less 'risky' shoes. Buy only what you need, watch sizes, do not go overboard on new fads.

  m Buying more staple merchandise. Buying more from fewer lines. Sticking with better nationally advertised merchandise.

  n No successful method with the current style situation. Manufacturers are experimenting, the retailer takes the markdowns – cuts gross profit by about 3 per cent – keep your stock at lowest level without losing sales.

5 Define a small sample of class members, work associates or friends, and ask them to answer the following in a paragraph or two: 'What are your career aspirations for the next five years?' Use one of the four basic units of content analysis to analyse their responses. Describe your findings as frequencies for the unit of analysis selected.

6 What is the median of the distribution 123, 154, 160, 187?

7 What happens to the mean and median in a set of five scores when the largest one is increased by several points? Set A: 12, 13, 23, 32, 43; Set A altered: 12, 13, 23, 32, 143.

From concept to practice

8 Either enter the values of Exhibit 16.8 into a spreadsheet or use the file data_ex16_8 provided on the website.

  a Compute their means, medians and modes. Which variables have similar means and medians?

  b Compute the standard deviations. Can you find any variables with proportionately larger standard deviations? What can you infer about the distributions' shapes?

  c Now compute skewness and kurtosis. Which variables have the least skewness? The most kurtosis?

9 On the website you find a data file named export_data and a codebook. The data are based on a survey among 60 firms. In the survey, those firms were asked about their export activities, alliance formation activities and their strategy.

  a Please run summary statistics and check how many observations you have for each variable.

  b Some of the variables have less than 60 observations. How can that be?

  c In which variables could one replace a missing value with 0.

Recommended further reading

Aczel, Amir D. and Jayauel Sounderpandian, *Complete Business Statistics* (7th edn). Chicago IL: McGraw-Hill, 2009. See Chapter 1 on descriptive statistics and Chapter 4 on the normal distribution.

Bigwood, Sally and Melissa Spore, *Presenting Numbers, Tables and Charts*. Oxford: Oxford University Press, 2003. A good text on how to present numbers.

Bux, William E. and Kenneth L. Gorman, *Data Entry Activities for Windows*. Cincinnati, OH: Thomson Learning, 2000. Fundamentals of data entry in a self-paced learning package suitable for novice computer users.

Pallant, Julie, *SPSS Survival Manual* (4th edn.). Open University Press, 2010. The bestseller on all guides explaining SPSS. Chapters 1–5 deal with all preparations and Chapters 6 and 7 explain how to calculate the descriptive statistics.

Strauss, Anselm and Juliet Corbin, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd edn). Thousand Oaks, CA: Sage Publishing, 2008. The book illustrates several approaches to coding procedures for qualitative data.

Get started with understanding statistical techniques!

When you have read this chapter, log on to the Online Learning Centre website at *www.mcgraw-hill.co.uk/textbooks/blumberg* to explore chapter-by-chapter test questions, additional case studies, a glossary and more online study tools for *Business Research Methods*.

Notes

1 Hans Zeisel, *Say It with Figures* (6th edn). New York: Harper & Row, 1985, pp. 48–9.

2 Jean M. Converse and Stanley Presser, *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage Publications, 1986, pp. 34–5.

3 Based on the operation of the SPSS, Inc. product TextSmart.

4 See the description of Remark Office OMR® found at: www.principiaproducts.com.

5 Remark Web Survey® is a product of Principia Products, Inc., 16 Industrial Blvd, Suite 102, Paoli, PA, 19301.

6 Adapted from a history of bar code development: www.lascofittings.com/BarCode-EDI/bchistory.htm.

7 For a thorough discussion of missing data analysis and remedies in multivariate data, see Joseph F. Hair Jr. et al., *Multivariate Data Analysis* (5th edn). Upper Saddle River, NJ: Prentice Hall, 1998, pp. 46–64.

8 Frederick Hartwig, with Brian E. Dearing, *Exploratory Data Analysis*. Beverly Hills, CA: Sage Publications, 1979, p. 15.